From:	Draper, Cynthia E
To:	Bondy, Garret E; Glover, Tim; Ellis, Steve; Curtis, Emmet F; Fortenberry, Chase; Griffith, Garry T.; jeff.keiser@CH2M.com; John Kern (kernstat@gmail.com) (kernstat@gmail.com); frank.dillon@ch2m.com; Bucholtz, Paul (DEQ); Saric, James; King, Todd W.; Canar, John; Milt Clark (mclark-59@comcast.net); Gendusa, Tony (GendusaTC@cdmsmith.com); Lavelle, James (LavelleJM@cdmsmith.com); Blischke, Eric (blischkee@cdmsmith.com); patricia.white@ch2m.com; Roth, Charles
Subject:	RE: Kalamazoo River OU5 Work Group
Date:	Thursday, April 18, 2013 11:59:41 AM
Attachments:	2013 04 18 Re-randomization White Paper.pdf

Please see attached re-randomization white paper for discussion on the call tomorrow.

-----Original Appointment-----
From: Draper, Cynthia E
Sent: Thursday, April 18, 2013 11:49 AM
To: Bondy, Garret E; Glover, Tim; Ellis, Steve; Curtis, Emmet F; 'Fortenberry, Chase'; 'Griffith, Garry T.'; 'jeff.keiser@CH2M.com'; 'John Kern (kernstat@gmail.com) (kernstat@gmail.com)'; frank.dillon@ch2m.com; 'Bucholtz, Paul (DEQ)'; 'Saric, James'; 'King, Todd W.'; 'Canar, John'; 'Milt Clark (mclark-59@comcast.net)'; 'Gendusa, Tony (GendusaTC@cdmsmith.com)'; 'Lavelle, James (LavelleJM@cdmsmith.com)'; 'Blischke, Eric (blischkee@cdmsmith.com)'; 'patricia.white@ch2m.com'; 'roth.charles@epa.gov'
Subject: Kalamazoo River OU5 Work Group
When: Friday, April 19, 2013 2:30 PM-4:30 PM (UTC-05:00) Eastern Time (US & Canada).
Where: 888-491-2632. 56636696

Work Group Conference Call, please see call in information below.

Call in number: 888- 491-2632
Code 56636696#

Agenda items:

1. Re-randomization technique and results of trial using example data
2. Discussion on basis for inclusion of samples into sample subset for Area 1 SWAC calculations
3. SWAC/Fish tissue trend path forward
4. Future call topics and scheduling

While we likely will not need 2 hours for this call, I have set aside that amount of time in the invite in case some individuals want to continue the discussion.

Documentation (white paper) for Agenda item 1 will be forwarded for your review today.

RE-RANDOMIZATION WHITE PAPER
GP KALAMAZOO, KALAMAZOO, MICHIGAN

1.0 PURPOSE

AMEC proposes that step-out samples can be used to calculate a SWAC using the re-randomization technique to randomly select from the available samples and come closer to estimating the "true" SWAC. Re-randomization of data is functionally similar to bootstrapping which is used by USEPA in statistical programs like ProUCL. It is possible that using re-randomization will remove some of the low bias of the step-out samples and allow incorporation of the non-random samples (i.e., step-out samples). Re-randomization will help reduce the high bias introduced from purposely selecting a greater percentage of fine-grained sediment sampling locations.

1.1 SWAC CONCEPT DEFINED

"Surface Weighted Average Concentration" or SWAC is a method of estimating the mean (average) concentration by randomly sampling an area, generating representative sub-areas for each sample location and generating a subarea-weighted average as the estimate. The sub-areas are often divided using such methods as Thiessen polygons and the stream tube method. These methods also have merit, but only the use of Thiessen polygons were used in this exercise at this time to demonstrate the re-randomization technique.

1.2 STEP-OUT SAMPLES' BIAS EFFECT

Additional step-out samples taken only around the highest value samples violates the random sampling assumption of the SWAC method and systematically biases the SWAC estimate downward by systematically "screening" the high value and reducing its representative sub-area's weight as discussed by Kern in the MDEQ comments (Kern). It should be noted that if step-out samples were collected around every sample location (not just the highest), then this systematic downward bias would disappear since all random points would be treated equivalently.

1.3 ADDITIONAL BIAS

An additional systematic bias has been identified in the actual real-world data set to be used for SWAC estimation. When selecting sampling locations along river transects during some sampling events, the subset of cores selected for PCB analysis was based on a desire to analyze more "fine-grained cores" and less "coarse-grained cores". This introduces a systematic bias against low concentration samples and introduces an additional upward bias to the SWAC estimations because PCBs are associated more strongly with fine-grained sediments than with coarse-grained sediments or gravels.

1.4 APPLICATION OF SWAC TO BIASED DATA

The resulting real world data set to be used in the SWAC procedure is double biased. The original data was taken as regular transect intervals, but was designed toreduced the amount of coarse sediment samples relative to the amount of fine sediment samples, resulting in a high bias. The step-out samples were taken only around high values biasing the data set lower. SWAC assumes true random samples for the weighted average method to generate a representative SWAC average value. What is needed is a method to transform the biased data into a true random (or at least approximately random) data set. One likely method is a re-randomization process which randomly re-samples from the existing dataset thus reducing the bias. The remainder of this document discusses the procedure and results of this re-randomization approach.

2.0 PROCEDURE

2.1 RE-RANDOMIZATION OVERVIEW

Re-sampling methods have a long history in the biological sciences, going back to at least the 1970s and the introduction of accessible general purpose digital computers. The theoretical basis goes back even further to Tukey in 1958 and others. The essential concept is taking an existing data set and taking random subsets from it, then performing a statistical calculation on that subset and recording the result. Then, the process (a “trial”) is repeated many times (hundreds or even thousands of times). The results of the many trials are then considered as a whole and the results are considered to estimate the statistic in question – often by taking the average of all trials or the median of the trials’ results.

The various re-randomization techniques vary in the process for choosing the subsamples for trials. In simple re-randomization, the subset is chosen without replacement, meaning no sample can be chosen more than once in a trial. Bootstrapping is very similar except it chooses with replacement – a sample may be chosen more than once for a trial. Both these methods produce acceptable results, but bootstrapping leads to numerical difficulties when dealing with spatially-located samples and area-weighted averages. The numerical difficulty is “What weight does a duplicate point have?”. Simple re-randomization without replacement discussed herein does not have this difficulty.

2.2 ESTIMATING POPULATION MEAN USING SWAC AND RE-RANDOMIZED SAMPLES

The statistic of interest in this exercise is the SWAC, so the statistic calculated in each trial is an area-weighted average using only the sample locations chosen in that trial.

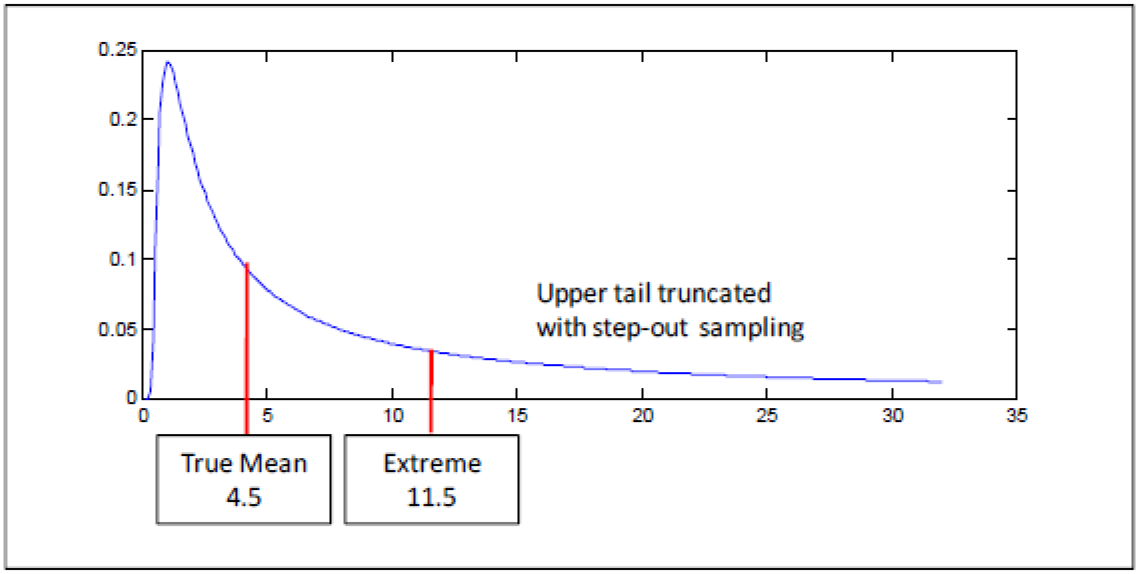
In Kern’s example, there are nine points on a regular grid and an additional four stepout samples taken around the highest value. The grid samples are randomly taken from a lognormal normal population with a mean (μ) of the natural logarithm (ln()) of the values being 1.0 and the standard deviation (σ) of the ln() values also being 1.0. It is not stated how the step-out sample values were selected. As stated in Kern, the true population mean of the population (not ln()) is $e^{\left(\mu+\frac{\sigma^2}{2}\right)}$ = 4.48. (Equation 1)

Table 1

x	y	value	group
5	5	2.1	original
5	10	3	original
5	15	3.7	original
7.5	10	3.3	stepout
10	12.5	1.2	stepout
10	5	1.9	original
10	7.5	3.4	stepout
10	15	5	original
10	10	11.5	original
12.5	10	0.9	stepout
15	15	1	original
15	5	6	original
15	10	6.9	original

2.3 Calculating the SWAC Weighted Average

The SWAC is generally determined as a simple arithmetic average of values weighted by their respective sub-areas. While the simple (or weighted) arithmetic average is considered a reasonable estimate of the population mean, for certain statistical distributions, it may be a biased estimate in some situations. A lognormal distribution with relatively few samples (like the current exercise) is one such situation. The simple or weighted average of a lognormal population with "few" samples tends to under-predict the true population mean. A review of the lognormal distribution graph suggests why this is the case.



(Kern Figure A-1)

Unlike the classic “Bell curve” the distribution is “pushed” to the left, making values less than the mean more likely than those greater than the mean. If “enough” values are taken, then the more common low values are balanced by the less-common (but much larger) higher values. However, this takes more than a “few” samples to balance out. So, having “few” samples (low-n) tends to under predict the true population mean.

How many is “few”? A general rule of thumb is less than 30, but there are exceptions. As an experiment, a number of synthetic data sets with varying numbers of members were taken from the described lognormal distribution and simple, un-weighted averages were calculated.

Table 2

number of values	mean
3	4.426
5	4.421
9	4.429
13	4.422
20	4.455
30	4.470
50	4.474
100	4.480

It appears at least 30 and perhaps more values are needed to closely approximate the true mean value of 4.48.

There is a less biased estimator of the mean as described in Gilbert (1987). It comes from Equation 1 above where the mean ($\bar{x}$) and the standard deviation (SD) of the actual sample data are substituted for the known population mean and standard deviation:

$$estimated\ mean = e^{\left(\bar{x}+\frac{SD^2}{2}\right)} \qquad \text{(Equation 2)}$$

The SWAC weighted average (of the ln() data) can be used in place of the mean. Pending the identification of a SWAC-like method of calculating the standard deviation, the un-weighted standard deviation of ln() is used as a close approximation.

3.0 RESULTS

Using the data set presented in Kern, a series of re-randomization trials were performed to generate an estimated SWAC-like concentration from the Kern data set. In each trial, 9 of the 13 data points were randomly chosen to make up a trial dataset. Using 5,000 re-randomization trials, the average SWAC-like estimated mean was 4.01. This estimated mean is larger than the original SWAC estimated mean using both the original and step-out data but less than the original SWAC mean using only the grid data (4.57). The estimated mean is also smaller than the straight, un-weighted average of just the grid data (4.57) but larger than the straight un-weighted average of both grid and step-out data (3.84) (Table 3).

Table 3

	Just Grid Data	Grid and Stepout
Original Kern SWAC Example	4.57	3.56
Kern un-weighted average	4.57	3.84
AMEC SWAC re-randomization	--	4.01
True Population Mean	4.48	4.48

The re-randomized SWAC mean estimate is somewhat less than the grid-only SWAC but higher than the grid-and-step-out SWAC estimate. The added step-out samples in the example are all lower values and less than the population mean – possibly a relic of how the step-out data was chosen – and may not be realistic. Therefore, performance of the re-randomization on real-life data may remove even more of the low-bias of step-out samples while allowing incorporation of non-random step-out samples.

AMEC ran 10 trials with the lognormal distribution used by Kern and randomly selected 4 values per trial to represent the "step-out" sample values (Table 4). Random selection was performed using the statistical package R. All resulting trials had at least one random value above the true mean, whereas Kern's step-out samples did not have any values greater than the true mean. This suggests a less than 1 in 10 chance that the step-out samples used in the Kern example are truly random. The use of any one of the these ten new trial "step-out" datasets would result in an estimated re-randomized SWAC greater than 4.01 which further reduces the alleged low bias. This suggests that the low bias might not be as extreme as initially suggested but may partially be an artifact of the "step-out" samples selected.

Table 4

Trial	Value 1	Value 2	Value 3	Value 4
1	0.55	4.80	2.15	1.66
2	0.59	17.17	2.46	1.42
3	1.83	2.32	7.41	16.21
4	2.60	5.63	0.46	14.39
5	2.57	0.53	6.51	1.99
6	5.28	0.42	0.64	2.22
7	1.57	6.02	0.39	1.36
8	13.69	2.14	4.99	2.62
9	8.90	1.97	1.27	2.95
10	1.99	3.83	3.22	10.28

The dataset used for evaluation was a synthetic dataset, not real transect data collected from the Kalamazoo River. Some of the transect data collection events selected a higher percentage of fine-grained sediment samples introducing a high bias in the dataset. This high bias is not present in this synthetic dataset and cannot be addressed or quantified in this exercise but this high bias in the real dataset will have the effect of overestimating the true SWAC within the Kalamazoo River.

4.0 SUMMARY

AMEC understands there is the concern that step-out samples have the potential to cause a low bias in the SWAC. AMEC worked through the example provided by Kern and found the following: 1) the use of a simple arithmetic average to calculate the SWAC when a less biased estimator of the mean would be more appropriate, 2) more samples for the synthetic data sample should be used to avoid under-predicting the true mean if a simple arithmetic average is to be used, and 3) the "selected" step-out samples were all below the true mean which is unlikely for a randomly selected set of points.

AMEC noted 1) that the real world samples were biased high by sampling, specifically because the focus was to analyze fine-grained cores during some sampling events and these locations would likely result in higher sample concentrations, and 2) that step-out sampling has a tendency to lower the concentration of samples since statistically a new sample would likely be lower than the original high concentration.

The re-randomization technique appears to reduce the downward bias from step-out samples, allows the incorporation of more numerous and more recent samples into the process while minimizing biases, and has the potential to reduce the high bias in real-world samples due to the core selection processes for some sampling events which focused PCB analysis on a higher percentage of fine-grained cores. AMEC proposes the use of the re-randomization technique for the generation of SWAC for the Kalamazoo River.